anales de psicología
2006, vol. 22, nº 1 (junio), 161-167

ISSN edición impresa: 0212-9728. ISSN edición web (www.um.es/analesps): 1695-2294

Procedimientos para detectar y medir el sesgo entre observadores*

Ana Benavente**, Manuel Ato y Juan J. López

Universidad de Murcia (España)

Resumen: En este trabajo se realiza un análisis de los distintos métodos para detectar y medir el sesgo entre observadores desde dos perspectivas básicas: el enfoque clásico, fundamentado en índices descriptivos y el enfoque del modelado, fundamentado en modelos loglineales. Se demuestra que estas medidas no son satisfactorias para detectar y medir el sesgo de forma unívoca porque presentan resultados contradictorios y se proponen nuevas alternativas a desarrollar que permitan descomponer correctamente error de medida y sesgo.
Palabras clave: Sesgo; acuerdo entre observadores; modelo log-lineal; modelo *mixture*.

Title: Methods for detecting and assessing the interrater bias.
Abstract: In this paper we present an analysis of the different methods commonly used to detect and assess interrater bias from two basic perspectives: classical approach based on descriptive-type criteria and loglinear model approach. We show that all these procedures are not satisfactory for the detection and measurement of observer bias in a univoque way due to contradictory results. We propose some new alternatives to develop which allow the correct separation of bias and measurement error.
Key words: Observer bias; rater agreement; log-linear model; mixture model.

A pesar de la amplia diversidad de procedimientos disponibles para el análisis del acuerdo entre observadores (*rater agreement*) con datos categóricos en las Ciencias Sociales y de la Salud (véase Uebersax, 2003), a través de la información bruta que aporta una tabla de acuerdo no es posible obtener medidas apropiadas que reflejen el grado de acuerdo real en presencia de heterogeneidad marginal y/o alta prevalencia en las categorías (Agresti, 2002), relacionados íntimamente con problemas tales como el error de medida de las variables y el sesgo entre observadores.

El efecto de sesgo de un observador respecto a otro ocurre cuando sus probabilidades marginales difieren, siendo mayor conforme aumenta la heterogeneidad de sus respectivas distribuciones marginales. En cierta medida vinculado al efecto de sesgo se encuentra el efecto de prevalencia, que ocurre en presencia de una proporción global extrema de resultados para una categoría de respuesta. En la práctica, representa la proporción de casos positivos de la población. Ambos efectos se han demostrado en varios trabajos (Spitznagel y Helzer, 1985; Feinstein y Cicchetti, 1990; Byrt, Bishop y Carlin 1993; Agresti, Ghosh y Bini, 1995; Lantz y Nebenzahl, 1996, y Hoehler, 2000).

El objetivo de este trabajo es analizar diferentes procedimientos para detectar y medir el sesgo entre observadores. Para ilustrar el problema utilizamos un ejemplo tomado de la investigación psicológica (Dillon y Mullani, 1984) que se analiza desde la perspectiva de los dos enfoques básicos citados (Ato, Benavente y López, en prensa).

1) *El enfoque descriptivo*, que se fundamenta en índices de tipo descriptivo (véase Zwick, 1988), el más popular de los cuales es el índice Kappa (κ) propuesto por Cohen (1960, 1968) para el caso de dos evaluadores, y su generalización para el caso de más de dos evaluadores (Fleiss, 1981). Kappa es un índice que se basa en el principio de corrección del azar *RCA* (*random corrected agreement*) cuya fórmula general es

$$RCA = \frac{(p_o - p_e)}{1 - p_e}$$

Todos los índices alternativos propuestos se basan en el mismo principio de corrección del azar, aunque utilizan fórmulas diferentes (Dunn, 1989; Shoukri, 2004).
Dos de las opciones más comunes que emplean el *RCA* se muestran en este trabajo, el índice κ (Cohen, 1960) y el índice π (Scott, 1955).

Para el caso del índice Kappa,

$$p_e = \sum p_{ei}^{\kappa} = \sum_{i=1}^{m} p_{i+} p_{+i} = \sum_{i=1}^{m} \left(\frac{n_{i+} n_{+i}}{N^2} \right)$$

$$p_o = \sum_{i} p_{oi}$$

El rango de valores del índice Kappa va de -1 a 1. La unidad representa el acuerdo perfecto, 0 indica que el acuerdo no es mejor que el acuerdo esperado por el azar, -1 implica acuerdo nulo.

Para el caso del índice π,

$$p_e^{\pi} = \sum p_{ei}^{\pi} = \sum_{i=1}^{m} \left(\frac{p_{i+} + p_{+i}}{2} \right)^2 = \sum_{i=1}^{m} \left(\frac{\frac{n_{i+}}{N} + \frac{n_{+i}}{N}}{2} \right)^2$$

Esta corrección asume que la distribución de las probabilidades marginales es homogénea para ambos observadores (supuesto de homogeneidad marginal).

El principio *RCA* es un procedimiento sencillo y universalmente aceptado para medir el acuerdo. No obstante, aunque los índices descriptivos han sido ampliamente utilizados en la literatura científica, especialmente en Ciencias del Comportamiento y en Ciencias Biológicas, muchos trabajos han puesto de manifiesto los problemas de estos índices, especialmente del índice Kappa (Brennan y Prediger,1981;

* **Nota**: Este trabajo ha sido financiado con fondos de un proyecto de investigación y desarrollo tecnológico concedido por el Ministerio de Educación y Ciencia (proyecto BSO2002-02513).
** **Dirección para correspondencia [Correspondence address]**: Ana Benavente Reche. Departamento de Psicología Básica y Metodología. Universidad de Murcia (Campus de Espinardo). Apartado 4031, 30080 Murcia (España). E-mail: anavent@um.es

Feinstein y Cichetti, 1990; Byrt, Bishop y Carlin, 1993; Agresti, Ghosh y Bini, 1995; Guggenmoos-Holtzman y Vonk, 1998; Nelson y Pepe, 2000), que muestra un comportamiento inadecuado en presencia de marginales heterogéneos y valores extremos de prevalencia de las categorías de respuesta. En general, dados valores iguales de p_o, cuanto más cercana a 0,5 sea la prevalencia, mayor será el valor de κ. Así, prevalencias muy bajas, o muy altas, penalizan el índice κ debido a que en ese caso la proporción de acuerdo esperado por azar es mayor que cuando la prevalencia es cercana a 0,5. Además, también se ve afectado por la simetría de los totales marginales. Cuanto mayor sea la diferencia entre los marginales mayor será el índice κ. Por otra parte, los índices de acuerdo descriptivos, y κ en particular, no permiten capturar los dos dimensiones básicas del acuerdo señaladas por Agresti (2002), la distintividad/asociación entre categorías y la ausencia de sesgo, ni separar la naturaleza del acuerdo y del desacuerdo,y se basan en algún modelo estadístico que en su aplicación práctica se asume como válido.

2) *El enfoque loglineal*, a diferencia de los índices clásicos, presenta la ventaja de descomponer el acuerdo observado en los dos componentes básicos: el acuerdo esperado por azar y el acuerdo no esperado por azar (Tanner y Young, 1985; Agresti, 1992). Otras ventajas adicionales que presentan los modelos loglineales son: (1) permiten probar el ajuste de los modelos así como la posibilidad de comparar una familia de modelos alternativos para encontrar el modelo óptimo, (2) utilizan el mismo rango de valores (de -1 a +1) de las medidas descriptivas, (3) permiten analizar pautas de acuerdo y desacuerdo entre dos o más observadores y compararlas cuando los sujetos se estratifican mediante una o más covariantes, y (4) muestran una gran flexibilidad que se generaliza también al desarrollo medidas de acuerdo (λ) de naturaleza similar a la de los índices estadísticos, aunque basados en una concepción distinta de la corrección del azar (Guggenmoss-Holtman y Vonk, 1998).

Este artículo se estructura como sigue. En la primera sección se introduce la notación y se presenta un ejemplo que se utilizará a lo largo del artículo, en la segunda se expone cómo detectar el sesgo utilizando los dos enfoques mostrados anteriormente, y en la tercera se analizan diferentes procedimientos para medir el sesgo. Finalmente se proponen algunas sugerencias alternativas que pueden resultar de interés para la investigación futura acerca de la detección y medición del sesgo entre observadores.

Notación y ejemplo

Supongamos que dos observadores A y B clasifican independientemente una muestra de *N* elementos (sujetos u objetos) en un mismo conjunto de *K* categorías nominales. El resultado de esta clasificación se puede resumir como se presenta en la Tabla 1, en la que n_{ij} representa las frecuencias observadas, p_{ij} representa las proporciones, *i* se refiere al observador A (fila) y *j* se refiere al observador B (columna), y n_{ij} representa el número de elementos que han sido clasificados por el observador A en la categoría *i* y por el observador B en la categoría *j*. La suma de las cantidades marginales de fila/columna produce el gran total (*N*).

Tabla 1: Notación general.

	Observador B						
Observador A	*1*	*2*	. . .	*j*	. . .	*K*	***Marginal A***
1	n_{11} p_{11}	n_{12} p_{12}	. . .	n_{1j} p_{1j}	. . .	n_{1K} p_{1K}	n_{1+} p_{1+}
2	n_{21} p_{21}	n_{22} p_{22}	. . .	n_{2j} p_{2j}	. . .	n_{2K} p_{2K}	n_{2+} p_{2+}
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
i	n_{i1} p_{i1}	n_{i2} p_{i2}	. . .	n_{ij} p_{ij}	. . .	n_{iK} p_{iK}	n_{i+} p_{i+}
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
K	n_{K1} p_{K1}	n_{K2} p_{K2}	. . .	n_{Kj} p_{Kj}	. . .	n_{KK} p_{KK}	n_{i+1} p_{i+}
Marginal B	n_{+1} p_{+1}	n_{+2} p_{+2}	. . .	n_{+j} p_{+j}	. . .	n_{+K} p_{+K}	$n_{++}=N$ $p_{++}=1$

Nota: n_{ij} representan frecuencias de respuesta; p_{ij} representan probabilidades de respuesta; p_{i+} y p_{+i} representan probabilidades marginales de fila y columna respectivamente.

En la Tabla 2 se muestra un ejemplo, tomado del trabajo de Dillon y Mullani, en el que dos observadores registraron un conjunto de 164 respuestas cognitivas elicitadas en un estudio de comunicación persuasiva sobre una escala con $K = 3$ categorías de respuesta ("positiva", "neutral" y "negativa").

Detección del sesgo

Entre los métodos más utilizados para detectar el sesgo entre observadores destacan aquellos que se basan en probar las hipótesis de homogeneidad marginal y simetría. La razón radica en la propia definición de sesgo. Como se definió previamente en la introducción, el sesgo de un observador se valora respecto de otro observador y se refiere a las discrepancias entre sus distribuciones marginales, por lo que aquel disminuye en la medida que las distribuciones marginales aproximan. La ausencia de sesgo implica que $p_{i+} = p_{+j}$ para todo *j* (Agresti, 1992).

Tabla 2: Frecuencias (y probabilidades) del ejemplo de Dillon y Mullani (1984).

Observador A	***Observador B***			
	Positiva	*Neutral*	*Negativa*	*Total*
Positiva	n_{11} **= 61** (0.37)	n_{12} **= 26** (0.16)	n_{13} **= 5** (0.03)	n_{1+} **= 92** (p_{1+} = 0.56)
Neutral	n_{21} **= 4** (0.02)	n_{22} **= 26** (0.16)	n_{23} **= 3** (f 0.02)	n_{2+} **= 33** (p_{2+} = 0.20)
Negativa	n_{31} **= 1** (0.01)	n_{32} **= 7** (0.04)	n_{33} **= 31** (0.19)	n_{3+} **= 39** (p_{3+} = 0.24)
Total	n_{+1} **= 66** (p_{+1} = 0.40)	n_{+1} **= 59** (p_{+2}= 0.36)	n_{+1} **= 39** (p_{+3} = 0.24)	n_{++}= N = **164** (p_{++} = 1.00)

Nota: Los números en negrilla son las frecuencias observadas y los números entre paréntesis
sus probabilidades.

La mayoría de los métodos para detectar el sesgo, al igual que los empleados para evaluarlo, se han adaptado para tablas de dimensión 2 x 2. En este trabajo se indican también algunos métodos al uso para tablas de mayor dimensión.

Procedimientos para detectar el sesgo en el enfoque clásico

En tablas de acuerdo, los evaluadores clasifican a los sujetos o ítems según una variable de interés con *K* niveles, de modo que para evaluar si las valoraciones de los observadores son iguales o divergentes se pueden aplicar pruebas que se basan en una distribución χ^2. Un resultado significativo implica que las frecuencias o probabilidades marginales no son homogéneas.

La prueba binomial exacta (Siegel y Castellan, 1988), se obtiene mediante

$$P1 = n_{12} / (n_{12} + n_{21}) \qquad \text{(Ec.1)}$$

cuya hipótesis nula permite probar si $P1 = .5$. La hipótesis de partida es H_o: $p_{1+} = p_{+1}$; H_1: $p_{1+} \neq p_{+1}$.

La prueba de McNemar (McNemar, 1947) para tablas 2 x 2 utiliza una distribución de χ_1^2 y se calcula aplicando

$$\chi_1^2 = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})} \qquad \text{(Ec.2a)}$$

Algunos autores recomiendan una versión de esta prueba con una corrección de la continuidad cuando los valores de n_{12} y/o n_{21} son pequeños (por ejemplo, $n_{12} + n_{21} < 10$), calculado como sigue:

$$\chi_1^2 = \frac{\left|(n_{12} - n_{21}) - 1\right|^2}{(n_{12} + n_{21})} \qquad \text{(Ec.2.b)}$$

Para tablas de acuerdo de mayor dimensión (*K* x *K*), una forma sencilla de calcular el sesgo a través de la prueba de McNemar, que describen Bishop, Fienberg y Holland (1975), consiste en aplicar la fórmula de la Ec. 1 con la diferencia de que n_{12} es igual a la suma de las frecuencias de las casillas del triángulo superior (las que se encuentran por encima de la diagonal principal), y n_{21} es igual a la suma de las frecuencias de las casillas del triángulo inferior (las que se encuentran por debajo de la diagonal principal). Así, por ejemplo, para una tabla 3 x 3 $n_{12} = n_{12} + n_{13} + n_{23}$ y $n_{21} = n_{21} + n_{31} + n_{32}$.

Con los datos de la Tabla 2 obtenemos χ_1^2 = 10.522 (P = .001). Puesto que resulta significativa se puede afirmar que no hay homogeneidad marginal.

La *extensión de Bowker* a la prueba de McNemar para una tabla cuadrada consiste en probar la hipótesis de simetría mediante H_o: $p_{ij} = p_{ji}$ y H_1: $p_{ij} \neq p_{ji}$. Se distribuye según $\chi^2_{K(K-1)/2}$ y viene dado por la siguiente ecuación (con corrección de la continuidad):

$$\chi^2_{K(K-1)/2} = \frac{\left|(n_{ij} - n_{ji}) - 1\right|^2}{(n_{ij} + n_{ji})} \qquad \text{(Ec.3)}$$

Al aplicar la Ec. 3 a los datos de la Tabla 2 observamos que χ_3^2 = 18.283 (P = .000), y por tanto cabe concluir existencia de sesgo.

El test de Stuart - Maxwell (Stuart, 1955; Maxwell, 1961, y Everitt, 1992) prueba si existe homogeneidad marginal en tablas de dimensión *K* x *K* para todas las categorías de forma simultánea. Se interpreta como una χ^2 con *K* - 1 grados de libertad. Para tablas 2 x 2, los resultados obtenidos con la prueba de Stuart - Maxwell y la prueba de McNemar son idénticos. El cálculo es algo complejo, basado en álgebra de matrices, pero puede obtenerse una aproximación para tablas de acuerdo pequeñas. Por ejemplo, para una tabla 3 x 3 la aproximación es la siguiente (Everitt, 1992):

$$\chi^2_{K-1} = \frac{\overline{n}_{23} d_1^2 + \overline{n}_{13} d_2^2 + \overline{n}_{12} d_3^2}{2(\overline{n}_{12}\overline{n}_{23} + \overline{n}_{12}\overline{n}_{13} + \overline{n}_{13}\overline{n}_{23})} \qquad \text{(Ec.4)}$$

donde $\overline{n}_{ij}$ es el promedio de las casillas simétricas; $n_{ij} = ½\left(n_{ij} + n_{ji}\right)$ y $d_i = \left(n_{i+} - n_{+1}\right)$, es decir, la diferencia entre los marginales de fila y columna.
Para los datos de la Tabla 2, la prueba Stuart-Maxwell adopta una valor de χ_2^2 = 45.066 (P = .000). Se concluye por tanto que no existe homogeneidad marginal.

Procedimientos para detectar el sesgo basados en el modelado estadístico loglineal

Los procedimientos basados en el enfoque clásico utilizan pruebas estadísticas de hipótesis nula mediante el contraste de hipótesis estadísticas. Una de las alternativas actualmente más consistentes es el enfoque del modelado estadístico, donde el concepto de modelo pasa a desempeñar un papel primordial (Ato *y otros*, 2005). Para detectar el sesgo mediante modelado estadístico se aplican modelos loglineales para probar si el modelo de *homogeneidad marginal* es consistente con los datos empíricos, es decir, que para cualquier tabla cuadrada se cumple:

$$p_i = \sum_{j=1}^{I} p_{ij} = \sum_{j=1}^{I} p_{ji} = p_{\bullet i}$$

Si, además, se trata de una tabla 2 x 2, entonces los modelos de simetría y homogeneidad marginal serán equivalentes. Sin embargo, para el caso general de tablas I x I debe tenerse presente que la simetría implica homogeneidad marginal pero que la homogeneidad marginal no implica necesariamente simetría.
Hay dos formas básicas de probar el modelo de homogeneidad marginal, una indirecta y otra directa. La forma indirecta se basa en una estrategia de ajuste condicional, donde asumiendo que se cumple el modelo de cuasi-simetría, el modelo de homogeneidad marginal es equivalente al modelo de simetría (Causinus, 1965):

Cuasi-Simetría (QS) + Homogeneidad marginal (HM) = Simetría (S) (Ec.5)

Al aplicar la Ec. 5 a los datos de la Tabla 2 se obtiene para el modelo de simetría una desvianza de L^2 (5) = 22.585 (P = .000), y para el modelo de cuasi-simetría una desvianza de L^2 (7) = 0.182 (P = .669). Puesto que el modelo de cuasi-simetría se ajusta óptimamente, la diferencia entre los modelos de simetría y cuasi-simetría nos proporciona una prueba aproximada del modelo de homogeneidad marginal, que en nuestro caso alcanza una desvianza $L^2(2)$ = 22.403 (P = .000), lo que evidencia un alto grado de desajuste. Por tanto, se concluye que los datos empíricos no son congruentes con el modelo de homogeneidad marginal.

Una forma directa de probar el modelo de homogeneidad marginal es a través del ajuste directo del modelo. Esta prueba es compleja, puesto que conlleva la aplicación de modelos marginales (véase Bergsma, 1998; Vermunt, Rodrigo y Ato, 2001), pero puede obtenerse utilizando una versión experimental del programa *LEM* (Vermunt, 1997). Tras aplicarlo a los datos del ejemplo se obtiene una desvianza de $L^2(2)$ = 22.081 (P = .000), lo que conduce a la misma conclusión anterior de que los datos empíricos no son consistentes con el modelo.

Medición del sesgo

Como sucede en la detección del sesgo, en su medición la mayoría de los procedimientos desarrollados hasta el momento son para tablas 2 x 2. Los índices empleados para evaluar el sesgo ofrecen un valor concreto pero no informan acerca de si son o no significativos. Estos índices se han desarrollado sólo para el enfoque descriptivo.

El índice de simetría en el desacuerdo (*symmetry in disagreement index*) para tablas 2 x 2 (Lanz y Nebenzahl, 1996) se calcula del modo siguiente:

$$S_D = (n_{12} - n_{21}) / (n_{12} + n_{21}) \quad \text{(Ec.6)}$$

que se distribuye según χ^2 con un grado de libertad. Adopta valores desde -1 a +1.

El índice de sesgo o BI (*bias index*) para tablas 2 x 2 (Byrt, Bishop y Carlin, 1993) presenta sin embargo un rango de valores que va de 0 a +1 y se obtiene mediante:

$$BI = | n_{12} - n_{21} | / N \quad \text{(Ec.7)}$$

Ludbrook (2004) propone una forma de evaluar el índice BI basado en el uso de una prueba exacta no condicional sobre las diferencias entre proporciones, en la que la hipótesis nula es $H_0 = p_1 - p_2 = 0$. Las proporciones binomiales son n_{12} / N y n_{21} / N. Las frecuencias binomiales correspondientes son $n_{12} / (N - n_{12})$ y $n_{21} / (N - n_{21})$ respectivamente.

Ludbrook (2002) extendió la evaluación del índice BI a tablas de dimensión K>2 aplicando la Ec. 7, definiendo n_{12} como la suma de todas las casillas que hay por encima de la diagonal principal ($\sum TS$: sumatorio de todos los elementos del triángulo superior) y n_{21} como la suma de todos los elementos (frecuencias) que hay por debajo de la diagonal principal ($\sum TI$: suma de todos los elementos del triángulo inferior). Las frecuencias binomiales correspondientes son $\frac{\sum TS}{\left(N - \sum TS\right)}$ y $\frac{\sum TI}{\left(N - \sum TI\right)}$. Al aplicar la Ec. 7 a los datos de la Tabla 2 junto con el método propuesto por Ludbrook para tablas 3 x 3 (que se ha complementado con el procedimiento de comparación múltiple de Holm), obtene-

mos un *BI* = 0.134 y una *P* = .0008, lo que implica diferencias estadísticamente significativas entre proporciones y por tanto denota la existencia de sesgo entre observadores.

El índice PABAK (*Prevalence and Bias Adjusted Kappa*)**,** propuesto por Byrt, Bishop y Carlin (1993) da un valor de Kappa corregido de sesgo y prevalencia para tablas de dimensión 2 x 2. Esencialmente toma Kappa y calcula una Kappa "equivalente" con una proporción con 50/50 de prevalencia y ausencia de sesgo. El índice se obtiene aplicando la siguiente fórmula:

$$PABAK = 2P_o - 1 = \kappa\,(1 - PI^2 + BI^2) + PI^2 - BI^2 \qquad \text{(Ec.8)}$$

donde $BI = n_{12} - n_{21}$ y $PI = n_{11} - n_{22}$

Los valores del índice PABAK varían de -1 a +1, al igual que el índice Kappa. La diferencia entre los valores reportados por los índices Kappa y PABAK nos aporta un valor de sesgo (en este caso también controlando la prevalencia). Sin embargo, sólo se puede aplicar a tablas de dimensión 2 x 2.

Arstein y Poesio (2005) proponen para tablas de cualquier dimensión un índice de sesgo basado en la diferencia entre el acuerdo esperado por azar para el índice π (Scott, 1955) y el acuerdo esperado por azar para el índice κ (Cohen, 1960) tal como se muestra a continuación:

$$B = \sum_i p_{ei}^{\pi} - \sum_i p_{ei}^{\kappa} \qquad \text{(Ec.9)}$$

Al aplicar la Ec.9 a los datos de Dillon y Mullani (Tabla 2) se obtiene una diferencia de B = 0.0164.

En las Tablas 3 y 4 se presentan respectivamente un resumen de los procedimientos estadísticos más relevantes actualmente existentes para detectar y medir el sesgo entre observadores.

Tabla 3: Resumen de los procedimientos para detectar el sesgo entre observadores.

PROCEDIMIENTOS PARA DETECTAR EL SESGO ENTRE OBSERVADORES			
1) Enfoque clásico			
Tablas 2 x 2			
Prueba binomial exacta (Siegel y Castellan, 1988)	$P = n_{12} / (n_{12} + n_{21})$ $P = 0.5$ $H_o: p_{1+} = p_{+1}$; $H_1: p_{1+} \neq p_{+1}$.		
Prueba Mcnemar (McNemar, 1947)	$\chi_1^2 = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})}$		
Tablas K > 2			
Extensión Mcnemar 3 x3 (Bishop, Fienberg y Holland, 1975)	$\chi_1^2 = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})}$ $n_{12} = n_{12} + n_{13} + n_{23}$ $n_{21} = n_{21} + n_{31} + n_{32}$ $H_o: p_{1+} = p_{+1}$; $H_1: p_{1+} \neq p_{+1}$		χ_1^2 = 10.522; P = .0012.
Prueba de Bowker (1948)	$\chi^2_{\frac{K(K-1)}{2}} = \frac{\left\|(n_{ij} - n_{ji}) - 1\right\|^2}{(n_{ij} + n_{ji})}$ $H_o: p_{ij} = p_{ji}$, y $H_1: p_{ij} \neq p_{ji}$		χ_3^2 = 18.283; P = .0004
Prueba de Stuart–Maxwell (Stuart, 1955; Maxwell, 1973 y Everitt, 1992)	$\chi^2_{K-1} = \frac{\bar{n}d_1^2 + \bar{n}_{13}d_2^2 + \bar{n}_{12}d_3^2}{2(\bar{n}_{12}\bar{n}_{23} + \bar{n}_{12}\bar{n}_{13} + \bar{n}_{13}\bar{n}_{23})}$		χ_2^2 = 45.066, P = .000
2) Modelos loglineales			
Tablas K > 2			
Prueba indirecta de hipótesis de homogeneidad marginal (Causinus, 1965)	Homogeneidad marginal (HM) = Simetría (S) - Cuasi-Simetría (QS)	(Ec.5)	L^2 (2)=22.403; P = .000
Prueba directa de hipótesis de homogeneidad marginal	Homogeneidad marginal (HM)		L^2 (2)= 22.081; P = .000

Tabla 4: Resumen de los procedimientos para evaluar el sesgo entre observadores.

PROCEDIMIENTOS PARA EVALUAR EL SESGO ENTRE OBSERVADORES		
	2) Modelado estadístico	
Tablas 2 x 2		
Indice de simetría en el desacuerdo (Lanz y Nebenzahl, 1996)	$S_D = (n_{12} - n_{21}) / (n_{12} + n_{21})$ *Rango -1 a +1*	
Indice de sesgo (Byrt; Bishop y Carlin, 1993; Ludbrook, 2004)	$BI = \lvert n_{12} - n_{21} \rvert / N$ $H_0 = p_1 - p_2 = 0.$ Rango 0 a +1	
PABAK (Byrt, Bishop y Carlin, 1993)	$PABAK = 2P_o - 1 = \kappa (1 - PI^2 + BI^2) + PI^2 - BI^2$ *Rango -1 a +1*	
Tablas K > 2		**Cuadro 2**
Indice de sesgo (Ludbrook, 2004)	$BI = \lvert n_{12} - n_{21} \rvert / N$ $H_0 = p_1 - p_2 = 0.$ $p1 = \frac{\sum TS}{\left(N - \sum TS\right)}$ $p2 = \frac{\sum TI}{\left(N - \sum TI\right)}$ Rango 0 a +1	BI = 0.134; P = 0.00084
Indice de sesgo (Arstein y Poesio, 2005)	$B = \sum_i p_{ei}^{\pi} - \sum_i p_{ei}^{\kappa}$	B = 0.0164

Conclusiones

La revisión de los procedimientos más relevantes actualmente existentes para la detección y medición del sesgo entre observadores nos permite llegar a la conclusión de que en la actualidad los investigadores aplicados de las Ciencias del Comportamiento y de las Ciencias Sociales no disponen de herramientas satisfactorias para obtener estimaciones fiables e insesgadas del sesgo. Como se infiere de las Tablas 3 y 4, se han propuesto distintas alternativas para detectar y medir el sesgo, pero la mayoría de ellas se basan en los datos brutos de una tabla de acuerdo y aplican procedimientos estadísticos globales que permiten responder a hipótesis concretas, pero no abordan dos aspectos fundamentales que justifican la existencia del sesgo entre observadores, a saber, la descomposición del grado de acuerdo y desacuerdo entre observadores y la separación de sesgo y error de medida en componentes mutuamente independientes. El primer aspecto ha recibido mucha atención en años recientes (Schuster, 2002; Schuster y Smith, 2002; Martín y Femia; 2004, Ato, Benavente y López, en prensa), pero el segundo sigue siendo uno de los aspectos olvidados de la investigación aplicada.

Desde nuestro punto de vista, los procedimientos para detectar y evaluar el sesgo deberían plantearse desde la perspectiva de los modelos con mezcla de distribuciones (Ato, Benavente y López, en prensa). Estos modelos asumen que los objetos que se clasifican en una tabla de acuerdo se extraen de una población compuesta por una mezcla de dos (o más) subpoblaciones finitas, cada una de las cuales identifica un conglomerado de objetos homogéneos, por ejemplo la subpoblación de acuerdo sistemático y la subpoblación de acuerdo aleatorio y desacuerdo. La exploración del sesgo podría abordarse más adelante si se contemplan nuevos componentes latentes (por ejemplo, ampliando a más de dos el número de clases latentes) o, incluso mejor, si se ampliara el número de variables latentes de un modelo *mezclas*. Una primera variable latente podría distinguir entre acuerdo y desacuerdo y una segunda variable latente, entre componente sistemático y componente aleatorio. En este contexto sería más directo definir apropiadamente la naturaleza del sesgo y buscar nuevos procedimientos para su detección y medida.

Referencias

Agresti, A. (1989). An agreement model with kappa as parameter. *Statistics and Probability Letters, 7*, 271-273.

Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research, 1,* 201-218.

Agresti, A. (2002). *Categorical Data Analysis*. 2nd Edition. Hoboken, NJ: Wiley.

Agresti, A., Ghosh, A. y Bini, M. (1995). Raking kappa: Describing potential impact of marginal distributions on measure of agreement, *Biometrical Journal, 37*, 811-820.

Arstein, R. y Poesio, M. (2005). *Kappa = Alpha (or Beta)*. CS Technical Report CSM-437. Essex, UK: University of Essex.

Ato, M., Benavente, A., Rabadán, R. y López, J.J. (2004). Modelos con mezcla de distribuciones para evaluar el acuerdo entre observadores. *Metodología de las Ciencias del Comportamiento, V. Especial 2004,* 47-54.

Ato, M., Benavente, A. y López, J.J. (en prensa). Análisis comparativo de tres enfoques para evaluar el acuerdo entre observadores. *Psicothema, 41.*

Bergsma, W. (1997). *Marginal Models for Categorical Data*. Tilburg, the Netherlands: Tilburg University Press.

Bishop, Y.M.M., Fienberg, S.E. y Holland, P.W. (1975). *Discrete Multivariate Analysis*. Cambridge, MA: The MIT Press.

Bowker, A.H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association, 43,* 572-574.

Brennan, R.L. y Prediger, D. (1981). Coefficient kappa: somes uses, misuses and alternatives. *Educational and Psychological Measurement, 41,* 687-699.

Byrt, T.; Bishop, J. y Carlin, J.B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46*, 423-429.

Causinus, H. (1965). Contribution à l'analyse statistique des tableaux de corrélation. *Annals Faculté of Sciences University of Toulouse, 29,* 77-182.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen. J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213-220.

Dillon, W.R. y Mullani, N. (1984). A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioral Research, 19*, 438-458.

Dunn, C. (1989). Design and Analysis of Reliability Studies: the statistical evaluation of measurement errors. Cambrigde, UK: Cambridge University Press.

Everitt, B.S. (1992). The Analysis of Contingency Tables. 2nd Edition. London, UK: Chapman and Hall.

Feinstein, A. y Cichetti, D. (1990). High agreement but low kappa: I. The problem of two paradoxes. *Journal of Clinical Epidemiology, 43*, 543-549.

Fleiss JL. (1981).Statistical methods for rates and proportions (second ed.) New York: Wiley.

Guggenmoos-Holtzmann, I. y Vonk, R. (1998). Kappa-like indices of observer agreement viewed from a latent class perspective. *Statistics in Medicine, 17*, 797-812.

Hoehler, F.K. (2000). Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology, 53*, 499-503.

Hsu, L.M. y Field, R. (2003). Interrater agreement measures: comments on kappa_n, Cohen's kappa, Scott's π and Aickin's α. *Understanding Statistics, 2*, 205-219.

Lantz, C.A. y Nebenzahl, E. (1996). Behavior and interpretation of the κ statistics: resolution of the two paradoxes. *Journal of Clinical Epidemiology, 49*, 431-434.

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: a critical review. *Clinical and Experimental Pharmacology and Physiology, 29,* 527-536.

Ludbrook, J. (2004). Detecting systematic bias between two raters. *Clinical and Experimental Pharmacology and Physiology, 31,* 113-115.

Martín, A. y Femia, P. (2004). Delta: a new measure of agreement between two raters. *British Journal of Mathematical and Statistical Psychology, 57*, 1-19.

Maxwell, A.E. (1961). *Analyzing Qualitative Data*. London, UK: Methuen.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12,* 153-157.

Nelson,J.C y Pepe,M.S. (2000). Statistical description of interrater variability in ordinal rating. *Statistical Methods in medical research, 5,* 475-496

Scott, W.A. (1955). Reliability of content analisys: The case of nominal scale coding. *Public Opinion, 19,* 321-325.

Schuster, C. (2002). A mixture model approach to indexing rater agreement. *British Journal of Mathematical and Statistical Psychology, 55*, 289-303.

Schuster, C. y Smith, D.A. (2002). Indexing systematic rater agreement with a latent-class model. *Psychological Methods, 7*, 384-395.

Shoukri, M.M. (2004). *Measures of Interobserver Agreement*. Boca Raton, Fl. CRC Press.

Siegel, S. y Castellan, N.J. (1988). Non parametric Statistics for the Behavioral Sciences. 2nd Edition. New York, NY: McGraw Hill.

Spitznagel, E.I. y Helzer, J.E. (1985). A proposed solution to the base rate problem in the kappa statistics. *Archives of General Psychiatry, 42,* 725-728.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika, 40*, 105-110.

Tanner, M.A. y Young, M.A. (1985a). Modeling ordinal scale disagreement. *Psychological Bulletin*, 98, 408-415.

Tanner, M.A. y Young, M.A. (1985b). Modeling agreement among raters. *Journal of the American Psychological Association, 80,* 175-180.

Vermunt, J.K. (1997). *LEM: a general program for the analysis of categorical data*. Tilburg: University of Tilburg.

Vermunt, J., Rodrigo, M.F. y Ato, M. (2001). Modeling joint and marginal distributions in the analysis of categorical panel data. *Sociological Methods and Research, 30*(2), 170-196.

Uebersax, J.S. (2003). *Statistical Methods for Rater Agreement*. Document download from: http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm

Von Eye, A. y Mun, E.Y. (2005). *Analyzing Rater Agreement*. Mahwah, NJ: Lawrence Erlbaum Associates.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103,* 374-378.

(Artículo recibido: 15-3-06; aceptado: 25-4-06)